Concept Article

Limitations and Uncertainties of Acute Fish Toxicity Assessments Can Be Reduced Using Alternative Methods

Martin Paparella[1#], Stefan Scholz[2#], Scott Belanger[3], Thomas Braunbeck[4], Pascal Bicherel[5], Kristin Connors[3], Christopher Faßbender[6], Marlies Halder[7], Adam Lillicrap[8], Roman Liska[7], Kristin Schirmer[9,10,11], Gilly Stoddart[6], Paul Thomas[5] and Susanne Walter-Rohde[12]

[1]Division of Medical Biochemistry, Biocenter, Medical University of Innsbruck, Innsbruck, Austria; [2]UFZ – Helmholtz Centre for Environmental Research, Department of Bioanalytical Ecotoxicology, Leipzig, Germany; [3]The Procter and Gamble Company, Cincinnati, OH, USA; [4]Aquatic Ecology & Toxicology, University of Heidelberg, INF 504, Heidelberg, Germany; [5]Knowledge & Research in Environment and Toxicology in Silico (KREATiS), L'Isle d'Abeau, France; [6]PETA International Science Consortium Ltd., London, United Kingdom; [7]European Commission, Joint Research Centre, Ispra, Italy; [8]Norwegian Institute for Water Research (NIVA), Oslo, Norway; [9]Eawag, Swiss Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland; [10]EPF Lausanne, School of Architecture, Civil and Environmental Engineering, Lausanne, Switzerland; [11]ETH Zürich, Department of Environmental Systems Science, Zürich, Switzerland; [12]German Environment Agency (UBA), Dessau-Roßlau, Germany

Abstract

Information about acute fish toxicity is routinely required in many jurisdictions for environmental risk assessment of chemicals. This information is typically obtained using a 96-hour juvenile fish test for lethality according to OECD test guideline (TG) 203 or equivalent regional guidelines. However, TG 203 has never been validated using the criteria currently required for new test methods including alternative methods. Characterization of the practicality and validity of TG 203 is important to provide a benchmark for alternative methods. This contribution systematically summarizes the available knowledge on limitations and uncertainties of TG 203, based on methodological, statistical, and biological considerations. Uncertainties stem from the historic flexibility (e.g., use of a broad range of species) and constraints of the basic test design (e.g., no replication). Other sources of uncertainty arise from environmental safety extrapolation based on TG 203 data. Environmental extrapolation models, combined with data from alternative methods, including mechanistic indicators of toxicity, may provide at least the same level of environmental protection. Yet, most importantly, the 3R advantages of alternative methods allow a better standardization, characterization, and an improved basic study design. This can enhance data reliability and thus facilitate the comparison of chemical toxicity, as well as the environmental classifications and prediction of no-effect concentrations of chemicals. Combined with the 3R gains and the potential for higher throughput, a reliable assessment of more chemicals can be achieved, leading to improved environmental protection.

1 Introduction

1.1 The current use of the *in vivo* acute fish toxicity test

The assessment of fish toxicity is an integral part of environmental hazard and risk assessment of many regulations worldwide (OECD, 2012a; Scholz et al., 2013). One of the frequently used vertebrate animal tests for aquatic toxicity assessments is the acute fish toxicity test (AFT)[1], which is typically conducted according to OECD Test Guideline 203 (TG 203) or similar guidelines (OECD, 2019b, 2012a; US EPA, 2016; ISO, 1996a,b,c). The AFT is used for the prospective assessment of individual chemicals, particularly to derive, depending on local

---

[1] For example, in Europe between 2015 and 2017, AFT testing caused about 45% to 60% of animal use from aquatic species for eco-toxicity testing, i.e., about 34,000 to 50,000 of 71,000 to 84,000 fish and amphibians, with amphibians contributing about 0.5% (European Commission, 2020, Part 2/5, tables 9.2 on pages 19, 53, 88)

The first two authors are ranked according to their contributions, thereafter alphabetically.

Disclaimer: The scientific views presented in this paper are those of the authors alone and do not necessarily reflect official views of their respective institutions.

Correspondence: Martin Paparella, PhD
Division of Medical Biochemistry, Biocenter
Medical University of Innsbruck
Innrain 80, 6020 Innsbruck, Austria
(martin.paparella@i-med.ac.at)

or international regulations, an environmental classification, a predicted no-effect concentration (PNEC) and/or one potential element of the toxicity criterion for PBT (persistence, bioaccumulation, toxicity) assessment (ECHA, 2017a). Furthermore, in some countries, the AFT is also conducted for effluent testing (Norberg-King et al., 2018; Scholz et al., 2013) or to inform the use of the test concentration for the fish bioconcentration test or as range finder for many other tests with more specific endpoints (OECD, 2012b). The AFT is based on a 96-h acute exposure of juvenile fish, identified as such by length. The percentage of lethality observed at each concentration is used to calculate an LC_{50} (lethal concentration at which 50% of the animals die). For the standard full concentration-response test, at least 5 concentrations with a minimum of 7 fish per concentration are used, without replication, resulting in a minimum number of 42 animals per test compound or sample (OECD, 2019b).

According to an OECD review (OECD, 2012a), the earlier TG 203 design of 1992 lacked critical specifications for several experimental parameters including the test duration (it was "preferably" 96 hours), use of solvents and solvent controls, application of statistical methods, measurement of fish length, and selection of a test species or multiple species from the numerous recommended species (Tab. S1[2], Section 9). Consequently, in 2019, TG 203 was revised to include more specifications and some test adaptations, including (as far as possible) the need for a validated analytical method to document actual test concentrations. The update also refers to revised guidance on the appropriate use of solvents (OECD, 2019a). However, the revision has not broadly changed the basic test design. For example, lethality was not replaced by moribundity as the definitive endpoint, the number of fish used per concentration remained 7 as a minimum, and the number of recommended fish species increased further to 11 (Tab. S1[2], Section 9).

Moreover, due to the limited stringency of some specifications in the early test protocols, which also partially apply to the revised version (fish species/strain/age cohorts, water conditions, use of moribundity or lethality, Section 2.2), the available data are very heterogeneous (Section 2.4.1 and Tab. S1[2], Section 9.1; Braunbeck et al., 2020).

1.2 The current use of alternatives to the *in vivo* acute fish toxicity test

Two experimental alternative methods have been standardized, validated and included in the OECD Test guidelines programme: The fish embryo acute toxicity test (FET), has been adopted as TG 236 (OECD, 2013, 2011a, 2012c; Busquet et al., 2014). The fish gill cell line acute toxicity test using the rainbow trout (*Oncorhynchus mykiss*) RTgill-W1 cell line has been scientifically validated (Fischer et al., 2019; Tanneberger et al., 2013; Natsch et al., 2018; ISO, 2019) and was included in the OECD WNT workplan as Project 2.63 for the development of a regulatory OECD test guideline in 2019 (Tab. S1[2], Sections 8.2, 9.2, 10.2).

Furthermore, computational approaches are available to predict acute fish toxicity either as freeware, such as US EPA TEST[3] and VEGA (Benfenati et al., 2013)[4], or as commercial software, such as CATALOGIC[5] and iSafeRat[6] (Thomas et al., 2019). Moreover, similarity of chemical structures and/or *in vitro* data may be used to form chemical categories with the purpose of supporting read-across of existing experimental *in vivo* data within those categories. The OECD QSAR Toolbox[7] may support such assessments (OECD, 2007; Low et al., 2013). Work towards more automated "big data" approaches is also in progress (Helman et al., 2019; Luechtefeld et al., 2018).

These alternative methods can provide mechanistic indicators for sub-lethal toxicity:

– Endpoints in the FET (coagulation of fertilized eggs, lack of somite formation, lack of detachment of the tail-bud from the yolk sac, and lack of heartbeat) are mechanistic in that they provide more information than simply whether a fish is dead or alive. If needed, these endpoints could also be expanded to include other endpoints such as neurotoxicity (Stengel et al., 2018; Zindler et al., 2019; Kluver et al., 2015).

– Based on the hypothesis that acute fish toxicity is often caused by nonspecific modes of action, endpoints such as metabolic activity and cell- and lysosomal membrane integrity measured in the RTgill W1 cell line test (Fischer et al., 2019) could be considered a mechanistic key event, even without a fully characterized adverse outcome pathway (Volz et al., 2011)[8].

– Similar to sub-lethal effects in the FET, such effects on gill

---

[3] Toxicity Estimation Software Tool (TEST). US Environmental Protection Agency. https://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test (updated 22.08.2016; accessed 15.05.2020).

[4] Virtual models for property evaluation of chemicals. Vegahub. https://www.vegahub.eu/ (accessed 15.05.2020).

[5] Software. Oasis Laboratory of Mathematical Chemistry. http://oasis-lmc.org/products/software/catalogic.aspx (accessed 15.05. 2020).

[6] iSafeRat Online. KREATiS. https://isaferat.kreatis.eu/en/qsar.php (accessed 15.05.2020).

[7] The OECD QSAR Toolbox. Organisation for Economic Co-operation and Development. http://www.oecd.org/chemicalsafety/risk-assessment/oecd-qsar-toolbox.htm (updated 15.04.2020; accessed 15.05.2020).

[8] The key relevance of cytotoxicity as a mechanism for acute toxicity was also recognized for acute mammalian toxicity (Prieto et al., 2019; Vinken and Blaauboer, 2017).

Abbreviations
AFT, Acute Fish Toxicity Test; CTD, chemical toxicity distribution; CV, coefficient of variation; EC_{50}, effective concentration 50%; EcoTTC, ecological threshold of concern; FET, Fish Embryo Acute Toxicity Test; GHS, Globally Harmonized System; GIVIMP, Good In Vitro Methods Practice; IATA, integrated approach to testing and assessment; ICE, interspecies correlation estimation; ITS, integrated testing strategy; LC_{50}, lethal concentration 50%, MAD, OECD Mutual Acceptance of Data system; MechoA, mechanism of action; MoA, mode of action; NOAEC, no observed adverse effect concentration; PBT, persistent bioaccumulating toxic; PNEC, predicted no-effect concentration; QAAR, quantitative activity-activity relationship; QSAR, quantitative structure-activity relationship; SSD, species sensitivity distribution; TG, test guideline; VMG Eco, (OECD) Validation Management Group Ecotoxicology; WNT, Working Group of National Coordinators of the OECD Test Guidelines Programme; WoE, weight-of-evidence

cells increase the likelihood of environmental population level lethality. This hypothesis appears mechanistically plausible, because compromised or weakened fish are likely to die in the real-world environment due to predators, competitors and/or other environmental stressors (Section 2.3 and Tab. S1[2], Section 1.1). Appropriate data to support this hypothesis are not available for fish, but for algae and invertebrates (Knillmann et al., 2012; Zhao et al., 2020).

– Finally, computational methods may inform on aquatic toxicity and structural alerts for non-baseline compounds and provide mechanistic information (Bauer et al., 2018a,b; Thomas et al., 2019). Such results may be integrated with data from alternative experimental methods within an integrated approach to testing and assessment (IATA) or Bayesian network approaches (see next paragraphs).

Since 2005, only one test using fish embryos, the fish-egg test (ISO, 2016) used in Germany as part of the waste water dues law, has been implemented as a stand-alone replacement of an acute fish toxicity test (Bundesgesetzblatt, 2005; Norberg-King et al.,

2018). In contrast, in chemical regulation worldwide, none of the experimental or computational alternative methods have been fully accepted as a stand-alone replacement for TG 203. Some, such as the FET and computational approaches, are considered useful at least within a weight-of-evidence (WoE) approach (ECHA, 2017b). However, the WoE and read-across-based approaches, which combine multiple sources of information, may be of limited regulatory efficiency and are rarely used due to their high complexity and low standardization. They could lead to subjectivity of data selection and integration, possibly resulting in disagreement between experts and low assessment throughput. Therefore, work towards a quantitative WoE approach to replace TG 203 using Bayesian networks has been initiated (Lillicrap et al., 2020; Moe et al., 2020).

A testing strategy that does not replace TG 203 but reduces the number of fish required is the threshold approach for acute fish toxicity, and this was standardized at OECD level (OECD, 2010). In this approach, standard acute toxicity tests not involving the use of vertebrate animals are first conducted with daphnids (OECD,

Tab. 1: Terminology used within this manuscript[a]

Term	Explanation	Potential improvements by alternative methods used within IATAs[b]
Limitation	Known practical disadvantages, e.g., testing throughput	yes
Variability	Variability is due to (biology and/or test guideline caused) diversity. For example:	
	Experimental variability is intrinsic to biology and due to flexibility in the TG. For a given TG, variability cannot be reduced with further knowledge.[c]	yes
	Environmental variability is intrinsic to biology and environment. It cannot be reduced with further knowledge.	no
Uncertainty	Uncertainty is due to limited knowledge about a true value including its (biology and/or test guideline caused) variability. For example:	
	The LC_{50} confidence interval is broad if the LC_{50} is close to the border of the concentration range tested and the concentration-response slope is flat.	yes
	The absence of a validation study implies uncertainty (= limited knowledge) about robustness and experimental variability of a method.	yes
	Awareness of the high diversity of real environments indicates uncertainty about quantitative knowledge of environmental variability.	no
Complexity	Complexity stems from multi-causal effect-relationships, e.g., in hazard characterization for aquatic life, based on a WoE assessment (including, e.g., read-across, QSAR, animal test data from superseded TGs, new alternative methods data), which is the result of a series of decisions, including, e.g., data sources, data quality assignments for their selection, similarity measures for read-across, weight assigned to various types of data in relation to mechanistic knowledge.	no
Ambiguity	Uncertainty stemming from the plurality of scientifically legitimate viewpoints, e.g., resulting from the complexity of scientific assessments.	no, by alternative methods; yes, by IATA guidance[d]

[a] The table is adapted from Paparella et al. (2017); for concepts of variability and uncertainty see, e.g., EFSA Scientific Committee et al. (2018), and for concepts of complexity and ambiguity see, e.g., IRGC (2017). [b] For discussion, see Figure 1 and text in Section 2 of this manuscript. [c] Knowledge about variability may be used to improve/change the test design. However, this results in a new TG with its new variability. [d] IATA guidance reduces the ambiguity by rules agreed *a priori* to testing and assessment. For example, the IATA for eye damage/irritation prescribes to carry out a Weight of Evidence (WoE) assessment based on available data *a priori* to new testing, and the result of this WoE determines the use of either of three different sequences of *in vitro* tests, and also the results at each step within the sequence of tests determines the need for follow-up testing (OECD, 2018a).

2004) and algae (OECD, 2011b). Using appropriate negative controls, fish are then exposed to the lowest EC_{50} of these tests at a single concentration (the threshold concentration) or using a limit test (100 mg/L), whichever concentration is lower. A full TG 203 concentration-response test is only performed if toxicity is observed at the threshold concentration. Since daphnids and algae are frequently the most sensitive trophic levels and, therefore, drive environmental classifications and PNECs, the threshold approach results in a significant reduction in the number of fish required for regulatory purposes (Jeram et al., 2005; Hutchinson et al., 2003). Recently, the possibility of including the FET in the threshold approach was explored (Rawlings et al., 2019) to support a time and cost-efficient use of the new method, and to optimize 3Rs gains for predicting acute fish toxicity.

The development of an "Integrated Approach to Testing and Assessment (IATA) for acute fish toxicity" was included in the OECD WNT work plan in 2015 (WNT project 2.54). In principle, an IATA for acute fish toxicity might be constructed similarly to the IATAs for skin or eye irritation (OECD, 2014, 2018a): First, a WoE assessment of all available and relevant information is conducted. This may already lead to a conclusion or could inform the need for follow-up testing using an integrated testing strategy (ITS). The ITS may aim at estimating whether the LC_{50} for fish is lower than for daphnids and/or algae, and only if this is the case, a more in-depth estimate for the fish LC_{50} should be provided. Such an ITS could represent an alternative to the current threshold approach by starting with acute tests with algae and daphnids, followed by QSARs, fish cell lines, and/or fish embryos, and conditionally – and only as a last resort and if indicated by the available data – would TG 203 be conducted. Computational approaches for data integration, e.g., Bayesian networks (Lillicrap et al., 2020; Moe et al., 2020), could complement the IATA, remove subjectivity, and provide an output in terms of a probability for a result.

The development of such an IATA with low potential for ambiguity (see Tab. 1, last line) may be essential for practical regulatory use and predictable acceptance by all stakeholders.

1.3 Transparency of scientific uncertainty is essential for responsible decision-making

This argument was already provided elsewhere (Paparella et al., 2020), but it is repeated and adapted here for the specific context.

It is essential for responsible decision-making in the management of chemicals that the uncertainties in data and knowledge are transparently described. This is important, as risk assessment and decision-making are typically carried out by different regulatory units or bodies. Guidance and tools have been developed for transparent characterization of the uncertainty of chemical risk metrics, such as ratios between human exposure and human limit values (EFSA Scientific Committee et al., 2018; WHO and IPCS, 2018).

However, there is a need for a similarly transparent analysis of uncertainties of the performance metrics for testing methods within the validation process. This has recently gained recognition in the field of human health regulatory toxicology, where, specifically, the uncertainty characterization of standard *in vivo* reference methods is starting to promote the acceptance of alter-

native methods. This experience from human health toxicology could also support the acceptance of alternative approaches to acute fish toxicity testing and assessment.

As an example, information on the reproducibility of test guidelines for animal tests in the field of eye irritation/damage and skin sensitization sets limits for achievable correlations between data from alternative methods and the animal test-based reference methods (Adriaens et al., 2014; Barroso et al., 2017; Hoffmann et al., 2018). It was also analyzed how the experimental variability of acute rodent LD_{50} data translates into variability of Global Harmonized System (GHS) classification (Hoffmann et al., 2010). Later, it was highlighted that, from a scientific perspective, a borderline range between GHS potency categories should be established. Test results falling into this borderline range should be considered as uncertain due to limited reliability of any test result (Leontaridou et al., 2017; Dimitrov et al., 2016). A comparable finding regarding aquatic acute toxicity classification has already been identified by Rawlings et al. (2019).

A systematic summary of uncertainties of animal reference methods is useful also in cases where a fully quantitative uncertainty characterization is not possible, due to a lack of data and/ or the complexity thereof. It allows at least a semi-quantitative and qualitative comparison of the performance and uncertainties of both *in vivo* and alternative approaches. This could support a best-informed decision on the acceptability of the new methodology. Such work was conducted for the rodent-based carcinogenicity assessment (Paparella et al., 2017) and is ongoing in the field of rodent-based developmental neurotoxicity assessment (Paparella et al., 2020).

Recently, the OECD Validation Management Group Ecotoxicology (VMG Eco) discussed that the uncertainties associated with long-standing OECD tests such as the TG 203 *in vivo* AFT should be compiled (2018, unpublished recommendations for updates of the fish testing framework (OECD, 2012a)). There are already several studies that analyze TG 203 LC_{50} variability (Hrovat et al., 2009; Scholz et al., 2016; Belanger et al., 2013; Busquet et al., 2014; Braunbeck et al., 2020; see Section 2.2 and Tab. S1[2], Sections 9.1 and 10). However, a more-in-depth summary of the potential limitations and uncertainties in variability and in environmental extrapolation of TG 203 is still lacking. The purpose of the present manuscript is to provide such a summary, applying an approach that has been used previously for the 2-year rodent cancer bioassay-based carcinogenicity assessment (Paparella et al., 2017). This approach builds on the existing OECD IATA guidance document (OECD, 2016) and suggests using identical structures for the characterization of the current method and the alternative approaches, including their specific uncertainties. This approach will facilitate a comprehensive comparative assessment in qualitative and quantitative terms.

2 Limitations and uncertainties of TG 203 – perspectives for reduction by alternative method-based IATAs

The use of the AFT, as conducted according to TG 203, is characterized by a number of limitations and uncertainties that could
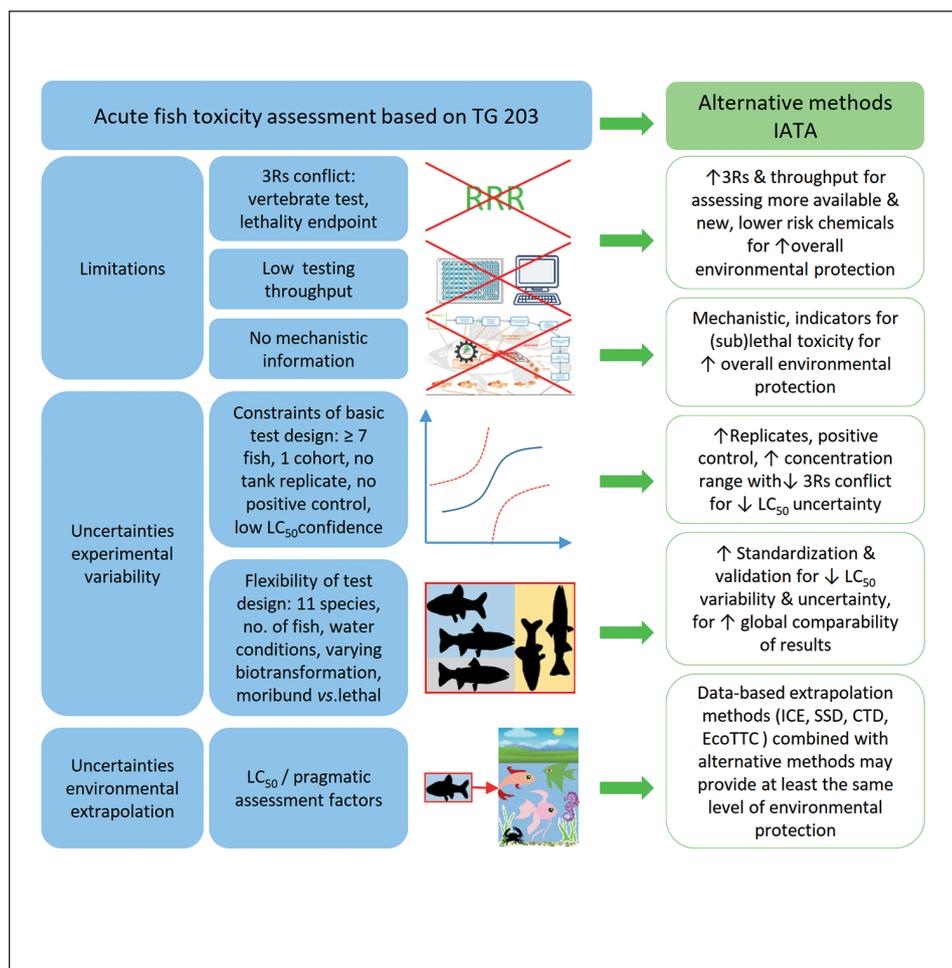
Fig. 1: Limitations and uncertainties of the OECD TG 203 acute fish toxicity test versus alternative methods-based IATAs
Acute fish toxicity assessment based on TG 203 contains limitations in terms of significant practical disadvantages as well as uncertainties in experimental variability, both of which may be reduced by the use of alternative methods combined within IATAs. Considering the uncertainties in real-world environmental extrapolations based on TG 203 data, data-based extrapolation methods combined with alternative methods may provide at least the same level of environmental protection. For further details, see the text in Section 2.1 to 2.3. ↑, increase; ↓, decrease; CTD, chemical toxicity distribution; EcoTTC, ecological threshold of concern; ICE, interspecies correlation estimation; SSD, species sensitivity distribution. (Sources of images: cell culture/computer & pond-image composition free from pixabay.com; fish-image for mechanism free from OECD AOP homepage; black fish drawings from Stefan Scholz)

be reduced by using alternative methods in the context of IATAs. For the presentation and discussion of these limitations and uncertainties, a similar approach was taken as published earlier for regulatory developmental neurotoxicity (Paparella et al., 2020): For a top-level overview, the main aspects of this discussion are illustrated in Figure 1 and further discussed in Section 2. Table 1 explains the terminology.

In Table S1[2], information about the limitations and uncertainties of TG 203 is presented within an OECD standard tabular format, which was originally developed to characterize alternative methods as individual information sources to be used within IATAs (OECD, 2016). This tabular summary was applied to carefully consider all potential limitations and uncertainties of the use of TG 203 and to develop the figure and text for Section 2 in this manuscript. Table S1[2] may be further amended and refined, as far as useful, in OECD VMG Eco. Applying the same systematic characterization scheme for both the use of TG 203 and alternative methods may support regulatory decision-making on the acceptability of the latter. The table could also be included in alternative methods-based IATA guidance documents as was the case for the standard *in vivo* study for the eye irritation/damage IATA (OECD, 2018a).

2.1 Limitations of TG 203 versus alternatives in terms of the 3Rs, testing-throughput and mechanistic information

Several limitations of TG 203 are inherent to the principle of this test and affect its practical regulatory applicability: This guideline is in direct conflict with the 3R goals (Russell and Burch, 1959), since it requires, for full concentration-response testing, at least 42 vertebrates of juvenile stages (a number causing statistical uncertainties, see Section 2.2), and this does not take into account the need for a range-finding study. Moreover, according to the TG, lethality shall be used as an endpoint, which further aggravates the concern. Termination of acute toxicity testing when moribundity is observed would result in improved animal welfare (Rufli, 2012), which is legally mandatory in Europe[8] and is currently applied in several countries. However, as explained below (Section 2.2, Paragraph 4), replacing lethality with moribundity as the endpoint was not agreed on globally at OECD level in the last update of TG 203.

The assay can only be conducted in a low-throughput manner. Rearing of fish is required to obtain the suitable juvenile size, which may take several weeks, depending on the species. Testing as such requires 96 hours, not including the additional time

for planning, preparation, assessment, reporting and tracking of culture health for approximately two weeks prior to testing. The throughput is further limited by the large volumes and vessels required to conduct the test. Therefore, it appears principally difficult to provide data for the more than 100,000 chemicals in commerce for which toxicity data are currently lacking[9] (Tab. S1[2], Section 16).

Moreover, the current endpoints in TG 203 provide little mechanistic information that could be useful for read-across and inferring chronic toxicity or supporting interspecies/environmental extrapolation modelling (cf. Section 2.3). Introducing moribundity and, perhaps, additional mechanistic endpoints by default into TG 203 would be theoretically feasible[10.] However, the current situation of limited quantity and quality of ecotoxicity data should be improved, e.g., according to the European strategy for a non-toxic environment[11] and the US vision for toxicity testing in the 21st century (NRC, 2007), and reaching this goal with traditional animal testing seems impossible.

In contrast, alternative approaches allow an increase in the testing throughput by using small-scale assessments and possibilities for automation. Alternative approaches would also allow testing of environmental degradation and reaction products (e.g., from disinfectants with biological material) and mixtures. Eventually, new chemicals that may have reduced environmental risk could be tested when available in laboratory-scale amounts only, and this may also promote the development of "green chemistry" (Maertens et al., 2014).

In summary, by providing more data for assessing many more chemicals and mixtures, alternative approaches may contribute to improved environmental safety without compromising global 3R goals (Tab. S1[2], Section 16.1).

2.2 Uncertainties in experimental variability relating to the study design of TG 203 versus alternatives

The basic study design of TG 203 causes uncertainty. Given the variability between individual fish and the use of a minimum of 7 fish per concentration from one cohort without tank replicates may lead to broad confidence intervals in LC_{50} estimates derived from concentration-response modelling, especially in the case of flat concentration-response relationships or when the LC_{50} is off-center relative to the boundaries of the tested concentration range (Tab. S1[2], Section 5.1; Carr et al., 2018). In addition, while the absence of study-internal positive controls is important to prevent further animal use, this causes uncertainty about potential intra- and inter-laboratory variability (Tab. S1[2], Section 4.3).

In contrast, alternative experimental methods allow an improved basic study design in terms of replicates, concentration ranges, and inclusion of study-internal positive controls. This is possible due to their small scale, potential for automation and 3Rs benefits. This improvement can reduce the uncertainty of results.

Some uncertainties of TG 203 relate to its level of standardization: For instance, TG 203 is flexible regarding the use of the test species. Any one of 11 recommended test species may be used, and guidance for selecting any of these is generic. According to TG 203, species selection should depend "*on regulatory requirements (industrial chemical, pharmaceutical, biocide or plant protection product, etc.) and on environmental exposure scenarios (cold, temperate or warm water species, freshwater or estuarine/marine fish)*" (OECD, 2019b). The possible use of diverse fish strains adds to this uncertainty. Other variables in the study design may also affect LC_{50} estimates, such as the test species-related water conditions (temperature, salinity, water hardness, pH). Also, the potentially variable age cohorts may affect the toxicity estimates (small differences within the recommended length range translate by cubic function to larger ranges of weight and developmental stage; cf. Tab. S1[2], Section 3 as well as Tab. 6 in Belanger et al., 2013). This diversity of potential test designs also means variability and uncertainty in variability of biotransformation in the AFT; little is known about species differences in this regard (cf. Tab. S1[2], Section 8; Braunbeck et al., 2020; Schlenk et al., 2008).

Furthermore, LC_{50} estimates in regulatory practice may be impacted by the inconsistent use of the endpoints lethality and moribundity (to conform with TG 203 and Directive 2010/63/EU, respectively[12]). On the one hand, the use of moribundity may reduce LC_{50} estimates on average by a factor of 2 (Rufli, 2012). On the other hand, observations of moribundity are likely more subjective than observations of mortality and may introduce additional variability to the assay result. However, the variability and uncertainty from the use of moribundity or lethality can be estimated and reduced as soon as unambiguous criteria for moribundity have been agreed upon (Tab. S1[2], Sections 1.1 and 4.1; Rufli, 2012). Thus, there is still uncertainty related to the use of these endpoints, but in principle the difference between lethality and moribundity could be scientifically calibrated (Tab. S1[2], Section 4.1).

An assessment of AFT LC_{50} values indicates a variability of up to a maximum range of 6 logarithmic units for the same chemical. However, this is based on historical data without application of stringent data quality filters (Hrovat et al., 2009). Two

---

[9] The estimate is based on the number of chemicals in the ECHA Classification and Labelling inventory, i.e., about 142,000 that should be on the European market. About 22,000 chemicals are registered for REACH at volumes of more than 1 ton per year. Acute toxicity studies for daphnids and algae are required for these. Only the about 7,000 substances registered above 10 tons per year require acute fish toxicity data. Slightly more than 60% of the data requirements were filled by experimental studies and the rest with read-across, QSAR or WoE assessments (ECHA, 2017c). Moreover, the available data are of heterogeneous quality (Braunbeck et al. 2020), *inter alia* due to a more limited standardization of the earlier versions of TG 203 (1981, 1983, 1992).

[10] As mentioned before, moribundity would also partly address the animal welfare concern. However, there is currently no agreement on the link between potential clinical signs and related mechanisms leading to mortality. Therefore, the OECD VMG-Eco agreed to incite voluntary collection of clinical signs first and their relation to lethality at a later stage. The use of moribundity would not improve the standardization or the throughput for testing and assessment. Moreover, it would require the development of reference databases and validation, which would necessitate to continue animal testing. Considering the current validation stage and perspectives for alternative methods, it may be more efficient to invest into alternative methods-based IATAs.

[11] European Commission (2017). http://ec.europa.eu/environment/chemicals/non-toxic/index_en.htm

[12] According to Directive 2010/63/EU, Article 13: "*Death as the end-point of a procedure shall be avoided as far as possible and replaced by early and humane end-points*".

older ring-trials indicate a maximum range of LC_{50} inter-laboratory variability of one logarithmic unit if fish interspecies variability is excluded but variability from other aspects such as fish size and exposure conditions (flow-through or static) is included. The two assessments were based on one chemical each, in one or two replicates and one or two fish species within 6 or 13 laboratories (Lemke, 1981; US EPA, 2001; Tab. S1[2], Section 10 also includes CVs for comparison with other ecotoxicity tests). However, a more comprehensive and most recent analysis applying stringent data quality filters indicated that about 8% and 0.5% of 181 chemicals showed differences in AFT LC_{50} data by factors of > 10 or > 100, respectively, when interspecies variability was excluded. When the TG 203 inherent interspecies variability was included, these percentages increased to about 15% and 10% of 53 chemicals (Braunbeck et al., 2020; Tab. S1[2], Section 10). Other work including AFT data for the neurotoxic biocide malathion indicates that AFT interspecies difference may be in the range of 4 orders of magnitude, depending on the chemical (Fig. 5 in Fischer et al., 2019). For compounds that require bioactivation, differences in the LC_{50} of > 50-fold have been identified for different fish species (Scholz et al., 2016). Overall, however, it is uncertain how the combination of all the variables within the TG 203 study design impact on the LC_{50} value (Tab. S1[2], Sections 9 and 10).

In principle, TG 203 could be standardized more stringently, similarly to TG 236 and more recent *in vitro* methods. This might reduce variability and uncertainty in LC_{50} values. However, at the OECD level, more standardization of TG 203 was not intended during the recent update process.

As theoretically indicated in the TG, the current flexibility could favor lethality estimates for a specific fish species and its specific environment or accommodate regulatory preferences. Moreover, the flexibility favors the practicability. Yet, increasing standardization of TG 203 now would not resolve the current heterogeneity of the historical database (Section 2.4.1) and could trigger a huge increase in regulatory demands for retesting with animals, which would conflict with 3R goals. It would also not provide the desired significant increase in overall global environmental protection, since the throughput would remain limited and the uncertainty related to the intention for more specific environmental extrapolations is underestimated (Sections 2.3 and 2.4.2).

Nonetheless, compared to TG 203, the experimental protocols of alternative methods already provide a higher level of standardization. This may reduce the variability of test results if the OECD guidance on Good In Vitro Method Practice (GIVIMP) is followed to develop a robust protocol (OECD, 2018b), like for the RTgill-W1 test (Tab. S1[2], Sections 3.2, 10 and 10.2). Moreover, comprehensive validation studies are available for alternative methods with well-standardized test protocols. Therefore, the uncertainty in the experimental variability estimates of test results may be relatively low (Tab. S1[2], Section 9.2). Low variability and, especially, low uncertainty in variability are significant advantages for environmental protection and global regulation in terms of PNEC calculation, GHS classification or the identification of the toxicity criterion for PBT assessment. Alternative methods may allow an increase in the global comparability of test results and thus reliably identify and globally regulate the – relative to all chemicals on the market – more toxic chemicals.

2.3 Uncertainties in environmental extrapolation – TG 203 versus alternatives

A relevant improvement for TG 203 would be to use moribundity instead of lethality as the endpoint. Lethality can represent a rather crude indicator for a chemical's potential to cause a population decrease in real environments (which is the ultimate intention of ecotoxicity testing). Moribundity might be an environmentally more relevant and protective endpoint. This hypothesis appears mechanistically plausible, since a weakened fish is likely to impact on populations in complex environmental situations, which include predators, competitors and/or other environmental stressors (Zhao et al., 2020; Knillmann et al., 2012; Rufli, 2012). However, as explained above (Section 2.2, Paragraph 4), this potential improvement has not been agreed globally at OECD level yet.

TG 203 is used in an attempt to assess the acute toxicity of a chemical to fish – based on the estimation of toxicity observed in only one developmental stage (juvenile) of one test species. LC/EC_{50} data from different organisms (fish, invertebrates, algae) are used in combination with pragmatic assessment factors to account for the potential variability in the sensitivity of different aquatic trophic levels (ECHA, 2008). Yet, the aquatic environment contains hundreds of thousands of species (Mora et al., 2011), various life-stages, and a vast array of abiotic and biotic modifiers. Data-based knowledge is available demonstrating that tests using a single species are "*in a majority of cases, reliable qualitative (some level of response seen) predictors of aquatic ecosystem community effects*" (de-Vlaming and Norberg-King, 1999). This latter US EPA review identified 57 studies (74%) that support this conclusion, 16 studies (21%) where single species testing underestimated aquatic ecosystem effects, and 4 studies (5%) that were inconclusive. The review also explains that full quantitative validation of single species tests through field studies is neither feasible nor meaningful given the huge environmental variability. This also means that a single environmentally "true" value does not exist, regardless of the assessment method applied (see below and Tab. S1[2], Section 2.1). However, the standard assessment factor approach results in a PNEC with an unknown level of environmental protection in terms of proportion of species under risk and related uncertainty. For GHS classification of chemicals, no assessment factors, but pragmatic cut-off values for the LC_{50} or EC_{50} from fish, daphnids and/or algae are used (if available), sometimes in combination with information on biodegradability and bioconcentration. This represents a similarly pragmatic approach (Tab. S1[2], Sections 6 and 7).

In summary, stand-alone LC_{50} values from TG 203 provide a rather uncertain basis for estimates of environmental toxicity (Tab. S1[2], Section 2).

Endpoints tested within alternative methods are not intended to be specific for any fish species, but rather a useful basis to estimate fish toxicity, at least for all current standard species in TG 203. Moreover, the endpoints may represent mechanistic indicators of an increased probability of fish population level lethali-

Tab. 2: Potential additional uncertainties from the use of alternative methods compared to similar uncertainties from current TG 203-based environmental safety assessment

Challenge	Alternative methods used for environmental safety assessment	*In vivo* AFT used for environmental safety assessment
Novel chemical with unknown/novel MOA	For completely new chemicals, which possess currently unknown modes of action, the relevance of the alternative method results may be more uncertain than for known chemical domains/MoAs. This is because such chemicals may not have been covered within the available environmental data analyses used for the generation and/or the performance assessment of alternative methods and computational models.	It is uncertain how many of the MoAs most relevant for environmental safety are covered with the AFT: The chemical domain of applicability has not been formally defined for the AFT. It is unlikely that the AFT, which does not cover effects on embryonic development, covers all possible environmentally relevant MoAs. Bearing in mind the hundreds of thousands of species and specific life-stages present in the aquatic environment (Mora et al., 2011), it is also uncertain which MoAs may be relevant for acute aquatic toxicity, but are not covered by testing usually only one species for each of the three trophic levels (fish, daphnids and algae).
Relevance to organism and population-level effects in the natural environment	Molecular/cellular level effects used as mechanistic *in vitro* indicators for potential toxicity may or may not be compensated at organism and/or population level. Such knowledge may remain limited due to the high complexity of mechanisms and interactions at organism, population and ecosystem level. (FET represents a test with an intact organism, thus uncertainty is limited to the extrapolation from the laboratory test to the environmental population and ecosystem level)	Sublethal effects and lethal effects at organism level in the laboratory may translate to various responses at population level in the different ecosystems, depending on the specific fauna and flora present, the specific food-webs, competitors and/or manifold other environmental variables and potential stressors (Zhao et al., 2020). Acute toxicity effects in the laboratory are "just" indicators for potential real-world environmental effects; see also discussion above (de-Vlaming and Norberg-King, 1999).
Biotransformation	*In vitro* biotransformation is limited to the biotransformation capacity of the isolated test system and may be dissimilar to *in vivo* biotransformation. Biotransformation in FET may differ from AFT.	Biotransformation may vary between species, life stages, and in response to environmental factors. Knowledge of this real-world environmental variability is very limited (Tab. S1[2]/Section 8).
Sensitivity across fish species	*In vitro* methods currently do not provide information on species-sensitivity differences. Theoretically FET may be carried out with different species. *In vitro* assays could also be created using cell lines from different fish species.	AFT can be conducted with a limited number of fish species that can be easily reared and/or tested in laboratory conditions. Furthermore, typically only one fish species is tested or required for regulation and hence, in regulatory practice, no chemical-specific species sensitivity comparison is conducted. Such testing does not include the variability of fish toxicity due to variable factors in the real-world environment. Moreover, fish may not be the most sensitive aquatic species. Hence, chemical-specific data about fish species variability may not significantly reduce uncertainty for environmental protection (Tab. S1[2], Sections 2.1. and 5 to 9).
	Sensitivity across fish species may not be a very critical concern for environmental protection: Since acute aquatic toxicity is a relatively data-rich field, we may not need more acute, fish species-specific data to improve available ICEs, SSDs, CTDs, EcoTTC for an effective environmental protection. Investigating relative sensitivities of further aquatic phyla for which there is less or no 3R conflict, like invertebrates and plants, might more effectively increase the environmental protection level.	

ty relative to other chemicals (Section 1.2 and Tab. S1[2], Section 1.2). Their mechanistic information content may also support inferring potential long-term impact as well as interspecies/environmental extrapolation modelling (see next paragraph). Considering the environmental extrapolation uncertainties associated with TG 203-derived PNECs and GHS classifications, alternative approaches may be expected to provide at least a similarly (un)certain or even improved environmental protection level. For example, differences between LC_{50} values derived from TG 203 and zebrafish embryo tests (e.g., OECD TG 236) have been found to be in the range of the fish species variability accepted within TG 203 (Lammer et al., 2009; Scholz et al., 2016).

Moreover, computational approaches are available for extrapolating experimental LC_{50} values to other environmental species. Interspecies correlation estimates (ICE), species sensitivity distributions (SSDs), chemical toxicity distributions (CTDs), and ecological threshold of concern (EcoTTCs) models (Bejarano et al., 2017; Belanger et al., 2015; Connors et al., 2019) provide estimates for a predicted species or trophic level effect without the need for additional animal test data. It is noted that the data used for model development stem from laboratory experiments and can never inform on the almost endless real environmental variability, but they allow best informed use of available knowledge (Tab. S1[2], Section 6.2). However, the development of these models and variability of model output might also be improved by less variable data from alternative methods.

It may be argued that the use of alternative methods introduces some additional and new uncertainties. Yet, these additional uncertainties may not significantly increase the current uncertainties for environmental safety assessment based on TG 203. Moreover, they may not be conceptually different (see Tab. 2).

In summary, a combination of alternative methods would allow the assessment of a larger number of chemicals, thereby promoting the identification and new development of chemicals with lower environmental risk. The improved basic test design and better standardization of alternative methods may also increase our ability to compare the toxicity of chemicals. In combination with environmental extrapolation models, data from alternative methods may provide mechanistic indicators of toxicity with at least the same level of environmental protection as current approaches.

2.4 Are further data needed to characterize the uncertainty of TG 203 data used in environmental hazard and risk assessment?

Uncertainties in experimental variability
TG 203 has never been formally validated. LC_{50} variability is uncertain, both for identical study designs and for all the study design variants covered within TG 203 (Section 2.2). It is not known how the variables within the TG 203 study design (selection of fish species, water conditions, life-stage, sex) affect the LC_{50}. If

we had a better understanding of how this variability in the test design can affect LC_{50} values, we would be able to calibrate any specific data set. Several reliability estimates for AFT data have been published, and these indicate a concern (Section 2.2, Tab. S1[2], Section 10). In order to better describe assay variability, a broader and more carefully curated AFT database could improve the quantitative assessment of variability (Braunbeck et al., 2020). However, such an extended retrospective assessment may be difficult or not feasible because robust historical data are generally scarce, i.e., about 15% of all available data (Braunbeck et al., 2020), and not available for all chemical groups and physical-chemical properties. Moreover, information on the variability of biotransformation in different fish species and life stages is scarce (Schlenk et al., 2008; Braunbeck et al., 2020). Therefore, it appears prudent to use the available data and other information to inform potential advances in regulatory science and decision-making for the acceptance of new, alternative approaches.

It should be considered at the regulatory science level whether it would be useful to develop a system using probabilistic (instead of deterministic) assignment of chemicals to the acute aquatic GHS category and the associated M-factors[13]. Categorization and M-factor attribution could be expressed in terms of a probability that the LC_{50} value is higher or lower than the acute category cut-off value of 1 mg/L and that the LC_{50} value will be within any of the 10-fold M-factor stratifications. In case the current variability of the standard threshold approach-based LC_{50} or EC_{50} (OECD, 2010) appears to be very high, refinement of the current GHS M-factor stratification might be considered. In view of the limitations, variability and uncertainties stemming from the use of TG 203 and the advantages of using alternatives, revision of the GHS classification criteria to explicitly include alternative approaches should be considered.

Uncertainties in environmental extrapolation
The need to use practically feasible approaches necessarily limits any type of testing and assessment in terms of its predictive capacity for the highly variable and complex aquatic environment. This, of course, also applies to the use of alternative approaches. It might be helpful to consider that any science-informed regulatory assessment needs to rely on some type of extrapolation model (Tab. S1[2], Section 6) – and this need indicates a conceptual similarity of animal tests and alternative tests.

Several tools, including ICE, SSD, CTD, and EcoTTCs models, can be used to describe, model, and account for the variability across species (Tab. S1[2], Sections 6.1-6.2). On this basis, a single experimental fish LC_{50} (or prediction thereof) can be recognized as a very uncertain estimate for environmental aquatic toxicity. Confidence intervals for more comprehensively informed environmental toxicity estimates, such as the 5th percentile of SSDs, may span 2 orders of magnitude (Bejarano et al., 2017; Awkerman et al., 2014). Moreover, the SSD models do not include the

---

[13] According to GHS and CLP regulation (EC No 1272/2008), LC_{50} values higher than 1 mg/L do not lead to acute aquatic toxicity classification, whereas LC_{50} values ≤ 1mg/L lead to classification into category 1. No other acute categories are defined, but further LC_{50} stratification into orders of magnitude (1-0.1 mg/L, 0.1-0.001 mg/L, etc.) allows the attribution of M-factors to the classified chemicals, and this supports better mixture classification by accounting for toxicity further to dilution of each constituent. Also, classification for chronic toxicity categories is possible based on acute fish LC_{50} data, depending on information for ready biodegradability and log Kow/BCF (ECHA, 2017c).

variability of fish toxicity due to the variability of environmental factors. However, it is important to acknowledge variabilities and uncertainties and review them with respect to the final regulatory use of the data. It should be recognized that any hazard characterization with standardized methods essentially represents a hazard comparison relative to other chemicals. Since in any case extrapolation uncertainties are huge, other aspects of scientific validity, such as biological/mechanistic relevance, low variability, and low uncertainty in variability, should receive increased attention. Alternative methods with reduced variability and uncertainty in variability (relative to the current TG 203) would at least better support the desired reliable hazard comparison between chemicals in terms of PNECs, GHS classifications, and identification of the toxicity criterion in the context of PBT assessment. It may be considered that alternative approach-based LC_{50} estimates could directly be used for environmental extrapolation model development (such as ICE, SSD, CTD, EcoTTc). Computational approaches for extrapolation to environmental relevance might even be improved by the use of less variable data from alternative methods. In summary, more knowledge to reduce the extrapolation uncertainty from TG 203 to the environment may not necessarily be needed for the immediate use and continued development of alternative methods.

2.5 Considering complexity for decision-making

Scientific data selection and knowledge integration is a complex task that may lead to a situation that different expert groups may come to different, scientifically legitimate conclusions. For example, PNECs derived by expert risk assessors on the basis of current standards can vary by 3 orders of magnitude, with the largest contributor being the heterogeneous judgment of study quality (Hahn et al., 2014). This represents ambiguity as one form of uncertainty, and it regularly appears within discussions on the regulatory validation and official acceptance of new alternative approaches. Validation is usually built on available complex data and information.

In situations of ambiguity, there may be a tendency to stick to the traditional approach. Typically, the existing animal test-based approaches for chemical hazard characterization are considered as the "gold standard". Given the experience and long-term use of the test, there is a high perceived confidence in the relevance of the result – albeit a robust validation of these tests is often lacking. In contrast, for alternative approaches, detailed validation studies have often been conducted, and the intra- and inter-laboratory variability is known, but perceived confidence is limited.

Imagine that an alternative approach represented the established standard assay and the AFT represented the new approach (Braunbeck et al., 2020): Are the available data for the AFT convincingly superior to the alternative in terms of uncertainty, variability, and environmental extrapolation? Would the available data be sufficient to replace the alternative method with the AFT?

In case of ambiguity, preference should be given to alternative approaches that can provide significant gains in terms of the 3Rs,

testing and assessment throughput, costs, and improved reliability. All these aspects are critically important to support testing and comparing the toxicity of many more chemicals in order to reduce the overall toxicological burden in the environment.

3 Conclusions

The current estimation of acute fish toxicity based on TG 203 bears various scientific uncertainties and practical regulatory limitations for achieving the final goal of protection of the environment from hazardous chemicals. The limitations relate to conflicts with the 3Rs principles, low throughput, and lack of mechanistic information. Uncertainties relate to experimental variability stemming from the basic study design and the study flexibility. Further uncertainties relate to the need for extrapolation to the highly variable environment.

Considering the interest in significantly improving the level of environmental protection, it is desirable that more reliable and comparable test data be generated and assessed for many more chemicals on the market as well as new chemicals intended to lower environmental risks. To achieve this aim, the future focus of regulatory toxicology needs to shift from individual WoE-based substance assessment towards development and harmonization of IATAs. These should be built on highly standardized alternative methods supported by computational approaches[14]. Such a strategy may be particularly important for lower-tier studies, which use relatively simple animal test guidelines with crude endpoints such as acute fish toxicity.

References

Adriaens, E., Barroso, J., Eskes, C. et al. (2014). Retrospective analysis of the Draize test for serious eye damage/eye irritation: Importance of understanding the in vivo endpoints under UN GHS/EU CLP for the development and evaluation of in vitro test methods. *Arch Toxicol 88*, 701-723. doi:10.1007/s00204-013-1156-8

Awkerman, J. A., Raimondo, S., Jackson, C. R. et al. (2014). Augmenting aquatic species sensitivity distributions with interspecies toxicity estimation models. *Environ Toxicol Chem 33*, 688-695. doi:10.1002/etc.2456

Barroso, J., Pfannenbecker, U., Adriaens, E. et al. (2017). Cosmetics Europe compilation of historical serious eye damage/eye irritation in vivo data analysed by drivers of classification to support the selection of chemicals for development and evaluation of alternative methods/strategies: The Draize eye test Reference Database (DRD). *Arch Toxicol 91*, 521-547. doi:10.1007/s00204-016-1679-x

Bauer, J. F., Thomas, P. C., Fouchard, S. Y. et al. (2018a). High-accuracy prediction of mechanisms of action using structural alerts. *Comput Toxicol 7*, 36-45. doi:10.1016/j.comtox.2018.06.004

---

[14] Establishing GLP-like quality control systems for computerized data assessment will also become an important aspect, and this is starting to be addressed at OECD level. Good Modelling Practice principles will support relevant model development and assessment EFSA (2014).

Bauer, J. F., Thomas, P. C., Fouchard, S. Y. et al. (2018b). A new classification algorithm based on mechanisms of action. *Comput Toxicol 5*, 8-15. doi:10.1016/j.comtox.2017.11.001

Bejarano, A. C., Raimondo, S. and Barron, M. G. (2017). Framework for optimizing selection of interspecies correlation estimation models to address species diversity and toxicity gaps in an aquatic database. *Environ Sci Technol 51*, 8158-8165. doi:10.1021/acs.est.7b01493

Belanger, S. E., Rawlings, J. M. and Carr, G. J. (2013). Use of fish embryo toxicity tests for the prediction of acute fish toxicity to chemicals. *Environ Toxicol Chem 32*, 1768-1783. doi:10.1002/etc.2244

Belanger, S. E., Sanderson, H., Embry, M. R. et al. (2015). It is time to develop ecological thresholds of toxicological concern to assist environmental hazard assessment. *Environ Toxicol Chem 34*, 2864-2869. doi:10.1002/etc.3132

Benfenati, E., Manganaro, A. and Gini, G. (2013). VEGA-QSAR: AI inside a platform for predictive toxicology. CEUR Workshop Proceedings. http://ceur-ws.org/Vol-1107/

Braunbeck, T., Böhler, S., Knörr, S. et al. (2020). Development of an OECD Guidance Document for the Application of OECD Test Guideline 236 (Acute Fish Embryo Toxicity Test): The chorion structure and biotransformation capacities of zebrafish as boundary conditions for OECD Test Guideline 236 – German contributions to OECD project 2.54: Guidance Document on IATA for Fish Acute Toxicity Testing. UBA Texte 94/2020.

Bundesgesetzblatt Bundesministerium für Umwelt, Naturschutz und Reaktorsicherheit (2005). Bekanntmachung der Neufassung des Abwasserabgabegesetzes. Teil I, Nr. 5, ausgegeben zu Bonn am 25 Januar 2005. http://extwprlegs1.fao.org/docs/pdf/ger35872b.pdf

Busquet, F., Strecker, R., Rawlings, J. M. et al. (2014). OECD validation study to assess intra- and inter-laboratory reproducibility of the zebrafish embryo toxicity test for acute aquatic toxicity testing. *Regul Toxicol Pharmacol 69*, 496-511. doi:10.1016/j.yrtph.2014.05.018

Carr, G. J., Bailer, A. J., Rawlings, J. M. et al. (2018). On the impact of sample size on median lethal concentration estimation in acute fish toxicity testing: Is n = 7/group enough? *Environ Toxicol Chem 37*, 1565-1578. doi:10.1002/etc.4098

Connors, K. A., Beasley, A., Barron, M. G. et al. (2019). Creation of a curated aquatic toxicology database: EnviroTox. *Environ Toxicol Chem 38*, 1062-1073. doi:10.1002/etc.4382

de-Vlaming, V. and Norberg-King, T. (1999). A review of single species toxicity tests: Are the tests reliable predictors of aquatic ecosystem community responses? EPA/600/R-97/114. https://bit.ly/3c1FI3T

Dimitrov, S., Detroyer, A., Piroird, C. et al. (2016). Accounting for data variability, a key factor in in vivo/in vitro relationships: Application to the skin sensitization potency (in vivo LLNA versus in vitro DPRA) example. *J Appl Toxicol 36*, 1568-1578. doi:10.1002/jat.3318

ECHA (2008). Guidance on Information Requirements and Chemical Safety Assessment. Chapter R.10: Characterisation of dose [concentration]-response for environment. Guidance for the implementation of REACH. https://bit.ly/37uWfg9

ECHA (2017a). Guidance on Information Requirements and Chemical Safety Assessment. Chapter R.11: PBT/vPvB assessment. Version 3.0. https://echa.europa.eu/guidance-documents/guidance-on-Information-requirements-and-chemical-safety-assessment

ECHA (2017b). Expert Workshop on the potential regulatory application of the Fish Embryo Acute Toxicity (FET) Test under REACH, CLP and the BPR, 3-4 May 2017 in Helsinki.

ECHA (2017c). Guidance on the Application of the CLP Criteria (Version 5.0). https://echa.europa.eu/guidance-documents/guidance-on-clp

EFSA (2014). Scientific opinion on good modelling practice in the context of mechanistic effect models for risk assessment of plant protection products. *EFSA J 12*, 3589. doi:10.2903/j.efsa.2014.3589

EFSA Scientific Committee, Benford, D., Halldorsson, T. et al. (2018). The principles and methods behind EFSA's guidance on uncertainty analysis in scientific assessment. *EFSA J 16*, 235. doi:10.2903/j.efsa.2018.5122

European Commission (2020). Report on the statistics on the number of animals used for experimental and other scientific purposes in the member states of the European Union in 2015-2017. Commission staff working document. https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1581689520921&uri=CELEX:52020SC0010

Fischer, M., Belanger, S. E., Berckmans, P. et al. (2019). Repeatability and reproducibility of the RTgill-W1 cell line assay for predicting fish acute toxicity. *Toxicol Sci 169*, 353-364. doi:10.1093/toxsci/kfz057

Hahn, T., Diamond, J., Dobson, S. et al. (2014). Predicted no effect concentration derivation as a significant source of variability in environmental hazard assessments of chemicals in aquatic systems: An international analysis. *Integr Environ Assess Manag 10*, 30-36. doi:10.1002/ieam.1473

Helman, G., Shah, I., Williams, A. J. et al. (2019). Generalized read-across (GenRA): A workflow implemented into the EPA CompTox chemicals dashboard. *ALTEX 36*, 462-465. doi:10.14573/altex.1811292

Hoffmann, S., Kinsner-Ovaskainen, A., Prieto, P. et al. (2010). Acute oral toxicity: Variability, reliability, relevance and interspecies comparison of rodent LD_{50} data from literature surveyed for the ACuteTox project. *Regul Toxicol Pharmacol 58*, 395-407. doi:10.1016/j.yrtph.2010.08.004

Hoffmann, S., Kleinstreuer, N., Alépée, N. et al. (2018). Non-animal methods to predict skin sensitization (I): The Cosmetics Europe database. *Crit Rev Toxicol 48*, 344-358. doi:10.1080/10408444.2018.1429385

Hrovat, M., Segner, H. and Jeram, S. (2009). Variability of in vivo fish acute toxicity data. *Regul Toxicol Pharmacol 54*, 294-300. doi:10.1016/j.yrtph.2009.05.013

Hutchinson, T. H., Barrett, S., Buzby, M. et al. (2003). A strategy to reduce the numbers of fish used in acute ecotoxicity testing of pharmaceuticals. *Environ Toxicol Chem 22*, 3031-3036. doi:10.1897/02-558

IRGC (2017). Introduction to the IRGC Risk Governance Framework, revised version. International Risk Governance Center. doi:10.1007/978-1-4020-6799-0_6

ISO – International-Standard-Organization (1996a). Water quali-

ty – Determination of the acute lethal toxicity of substances to a freshwater fish [Brachydanio rerio Hamilton-Buchanan (Teleostei, Cyprinidae)] – Part 3: Flow-through method. ISO 7346-3:1996. doi:10.3403/00151595u

ISO (1996b). Water quality – Determination of the acute lethal toxicity of substances to a freshwater fish [Brachydanio rerio Hamilton-Buchanan (Teleostei, Cyprinidae)] – Part 2: Semi-static method. ISO 7346-2:1996. doi:10.3403/00151583u

ISO (1996c). Water quality – Determination of the acute lethal toxicity of substances to a freshwater fish [Brachydanio rerio Hamilton-Buchanan (Teleostei, Cyprinidae)] – Part 1: Static method. ISO 7346-1:1996. doi:10.3403/00151583u

ISO (2016). Water quality – Determination of the acute toxicity of waste water to zebrafish eggs (Danio rerio). ISO 15088:2007. doi:10.31030/1495364

ISO (2019). Water quality – Determination of acute toxicity of water samples and chemicals to a fish gill cell line (RTgill-W1). ISO 21115:2019. doi:10.3403/30361171u

Jeram, S., Sintes, J. M., Halder, M. et al. (2005). A strategy to reduce the use of fish in acute ecotoxicity testing of new chemical substances notified in the European Union. *Regul Toxicol Pharmacol 42*, 218-224. doi:10.1016/j.yrtph.2005.04.005

Kluver, N., Konig, M., Ortmann, J. et al. (2015). Fish embryo toxicity test: Identification of compounds with weak toxicity and analysis of behavioral effects to improve prediction of acute toxicity for neurotoxic compounds. *Environ Sci Technol 49*, 7002-7011. doi:10.1021/acs.est.5b01910

Knillmann, S., Stampfli, N. C., Beketov, M. A. et al. (2012). Intraspecific competition increases toxicant effects in outdoor pond microcosms. *Ecotoxicology 21*, 1857-1866. doi:10.1007/s10646-012-0919-y

Lammer, E., Carr, G. J., Wendler, K. et al. (2009). Is the fish embryo toxicity test (FET) with the zebrafish (Danio rerio) a potential alternative for the fish acute toxicity test? *Comp Biochem Physiol C Toxicol Pharmacol 149*, 196-209. doi:10.1016/j.cbpc.2008.11.006

Lemke, A. (1981). Interlaboratory Comparision: Acute Testing Set. US-EPA EPA-600/S3-81-005. https://nepis.epa.gov/Exe/ZyPURL.cgi?Dockey=2000TSC7.TXT

Leontaridou, M., Urbisch, D., Kolle, S. N. et al. (2017). The borderline range of toxicological methods: Quantification and implications for evaluating precision. *ALTEX 34*, 525-538. doi:10.14573/altex.1606271

Lillicrap, A., Moe, S. J., Wolf, R. et al. (2020). Evaluation of a Bayesian network for predicting acute fish toxicity from fish embryo toxicity data. *Integr Environ Assess Manag 16*, 452-460. doi:10.1002/ieam.4258

Low, Y., Sedykh, A., Fourches, D. et al. (2013). Integrative chemical-biological read-across approach for chemical hazard classification. *Chem Res Toxicol 26*, 1199-1208. doi:10.1021/tx400110f

Luechtefeld, T., Marsh, D., Rowlands, C. et al. (2018). Machine learning of toxicological big data enables read-across structure activity relationships (RASAR) outperforming animal test reproducibility. *Toxicol Sci 165*, 198-212. doi:10.1093/toxsci/kfy152

Maertens, A., Anastas, N., Spencer, P. J. et al. (2014). Green toxicology. *ALTEX 31*, 243-249. doi:10.14573/altex.1406181

Moe, S. J., Madsen, A. L., Connors, K. A. et al. (2020).

Development of a hybrid Bayesian network model for predicting acute fish toxicity using multiple lines of evidence. *Environ Modell Softw 126*, 104655. doi:10.1016/j.envsoft.2020.104655

Mora, C., Tittensor, D. P., Adl, S. et al. (2011). How many species are there on Earth and in the ocean? *PLoS Biol 9*, e1001127. doi:10.1371/journal.pbio.1001127

NRC – National Research Council (2007). *Toxicity Testing in the 21st Century: A Vision and a Strategy*. Washington, DC, USA: The National Academies Press. doi:10.17226/11970

Natsch, A., Laue, H., Haupt, T. et al. (2018). Accurate prediction of acute fish toxicity of fragrance chemicals with the RTgill-W1 cell assay. *Environ Toxicol Chem 37*, 931-941. doi:10.1002/etc.4027

Norberg-King, T. J., Embry, M. R., Belanger, S. E. et al. (2018). An international perspective on the tools and concepts for effluent toxicity assessments in the context of animal alternatives: Reduction in vertebrate use. *Environ Toxicol Chem 37*, 2745-2757. doi:10.1002/etc.4259

OECD (2004). Test No. 202: Daphnia sp. Acute Immobilisation Test. *OECD Guidelines for the Testing of Chemicals, Section 2*. OECD Publishing, Paris. doi:10.1787/9789264069947-en

OECD (2007). Guidance on Grouping of Chemicals. *OECD Series on Testing and Assessment, No. 80*. OECD Publishing, Paris. https://bit.ly/37so20A

OECD (2010). Short Guidance on the Threshold Approach for Acute Fish Toxicity. *OECD Series of Testing and Assessment, No. 126*. OECD Publishing, Paris. https://bit.ly/3mrLsHF

OECD (2011a). Validation Report (Phase 1) for the Zebrafish Embryo Toxicity Test. *OECD Series of Testing and Assessment, No. 157*. OECD Publishing, Paris. http://www.oecd.org/env/ehs/testing/48572244.pdf

OECD (2011b). Test No. 201: Freshwater Alga and Cyanobacteria, Growth Inhibition Test. *OECD Guidelines for the Testing of Chemicals, Section 2*. OECD Publishing, Paris. doi:10.1787/9789264069923-en

OECD (2012a). Fish Testing Framework. *OECD Series of Testing and Assessment, No. 171*. OECD Publishing, Paris. https://bit.ly/3r37l3A

OECD (2012b). Test No. 305: Bioaccumulation in Fish: Aqueous and Dietary Exposure. *OECD Guidelines for the Testing of Chemicals*. OECD Publishing, Paris. doi:10.1787/9789264185296-en

OECD (2012c). Validation Report (Phase 2) for the Zebrafish Embryo Toxicity Test. *OECD Series of Testing and Assessment, No. 179*. OECD Publishing, Paris. https://bit.ly/2LFH7Ut

OECD (2013). Test No. 236: Fish Embryo Acute Toxicity (FET) Test. *OECD Guidelines for the Testing of Chemicals, Section 2*. OECD Publishing, Paris. doi:10.1787/9789264203709-en

OECD (2014). New Guidance Document on an Integrated Approach to Testing and Assessment (IATA) for Skin Corrosion and Irritation. *Series of Testing and Assessment, No. 203*. OECD Publishing, Paris. https://bit.ly/38cGj1d

OECD (2016). Guidance Document on the Reporting of Defined Approaches to be Used Within Integrated Approaches to Testing and Assessment. *OECD Series on Testing and Assessment, No. 255*. OECD Publishing, Paris. https://bit.ly/37qsEo2

OECD (2018a). Guidance Document on Integrated Approaches to Testing and Assessment (IATA) for Serious Eye Damage and Eye Irritation. *OECD Series on Testing and Assessment, No. 263*. OECD Publishing, Paris. doi:10.1787/84b83321-en

OECD (2018b). Guidance Document on Good In Vitro Method Practice (GIVIMP). *OECD Series of Testing and Assessment, No. 286*. OECD Publishing, Paris. doi:10.1787/9789264304796-21-en

OECD (2019a). Guidance Document on Aqueous-Phase Aquatic Toxicity Testing of Difficult Test Chemicals. *OECD Series of Testing and Assessment, No. 23* (second edition). OECD Publishing, Paris. doi:10.1787/0ed2f88e-en

OECD (2019b). Test Guideline No. 203: Fish, Acute Toxicity Testing. *OECD Test Guidelines for the Testing of Chemicals, Section 2*. OECD Publishing, Paris. doi:10.1787/9789264069961-en

Paparella, M., Colacci, A. and Jacobs, M. N. (2017). Uncertainties of testing methods: What do we (want to) know about carcinogenicity? *ALTEX 34*, 235-252. doi:10.14573/altex.1608281

Paparella, M., Bennekou, S. H. and Bal-Price, A. (2020). An analysis of the limitations and uncertainties of in vivo developmental neurotoxicity testing and assessment to identify the potential for alternative approaches. *Reprod Toxicol 96*, 327-336. doi:10.1016/j.reprotox.2020.08.002

Prieto, P., Graepel, R., Gerloff, K. et al. (2019). Investigating cell type specific mechanisms contributing to acute oral toxicity. *ALTEX 36*, 39-64. doi:10.14573/altex.1805181

Rawlings, J. M., Belanger, S. E., Connors, K. A. et al. (2019). Fish embryo tests and acute fish toxicity tests are interchangeable in the application of the threshold approach. *Environ Toxicol Chem 38*, 671-681. doi:10.1002/etc.4351

Rufli, H. (2012). Introduction of moribund category to OECD fish acute test and its effect on suffering and LC$_{50}$ values. *Environ Toxicol Chem 31*, 1107-1112. doi:10.1002/etc.1779

Russell, W. M. S. and Burch, R. L. (1959). *The Principles of Humane Experimental Technique*. https://caat.jhsph.edu/principles/the-principles-of-humane-experimental-technique

Schlenk, D., Celander, M., Gallagher, E. et al. (2008). Biotransformation in fishes. In R. T. Di Giulio and D. E. Hinton (eds), *The Toxicology of Fishes*. Boca Raton, USA: CRC Press. doi:10.1201/9780203647295

Scholz, S., Sela, E., Blaha, L. et al. (2013). A European perspective on alternatives to animal testing for environmental hazard identification and risk assessment. *Regul Toxicol Pharmacol 67*, 506-530. doi:10.1016/j.yrtph.2013.10.003

Scholz, S., Klüver, N. and Kühne, R. (2016). Analysis of the relevance and adequateness of using Fish Embryo Acute Toxicity (FET) Test Guidance (OECD 236) to fulfil the information requirements and addressing concerns under REACH. Report ECHA-UFZ contract ECHA/2014/341, 1-105.

Stengel, D., Wahby, S. and Braunbeck, T. (2018). In search of a comprehensible set of endpoints for the routine monitoring of neurotoxicity in vertebrates: Sensory perception and nerve transmission in zebrafish (Danio rerio) embryos. *Environ Sci Pollut Res Int 25*, 4066-4084. doi:10.1007/s11356-017-0399-y

Tanneberger, K., Knobel, M., Busser, F. J. et al. (2013). Predicting fish acute toxicity using a fish gill cell line-based toxicity assay. *Environ Sci Technol 47*, 1110-1119. doi:10.1021/es303505z

Thomas, P. C., Bicherel, P. and Bauer, F. J. (2019). How in silico and QSAR approaches can increase confidence in environmental hazard and risk assessment. *Integr Environ Assess Manag 15*, 40-50. doi:10.1002/ieam.4108

US EPA (2001). Final Report: Interlaboratory Variability Study of EPA Short-term Chronic and Acute Whole Effluent Toxicity Test Methods, Vol. 1. EPA 821-B-01-004. Office of Water, U.S. Environmental Protection Agency, Washington, D.C. https://bit.ly/3nraCaU

US EPA (2016). OCSPP 850.1075: Freshwater and Saltwater Fish Acute Toxicity Test. Ecological Effects Test Guidelines. https://www.epa.gov/test-guidelines-pesticides-and-toxic-substances/series-850-ecological-effects-test-guidelines

Vinken, M. and Blaauboer, B. J. (2017). In vitro testing of basal cytotoxicity: Establishment of an adverse outcome pathway from chemical insult to cell death. *Toxicol In Vitro 39*, 104-110. doi:10.1016/j.tiv.2016.12.004

Volz, D. C., Belanger, S., Embry, M. et al. (2011). Adverse outcome pathways during early fish development: A conceptual framework for identification of chemical screening and prioritization strategies. *Toxicol Sci 123*, 349-358. doi:10.1093/toxsci/kfr185

WHO – World Health Organization and IPCS – International Programme on Chemical Safety (2018). Guidance document on evaluating and expressing uncertainty in hazard characterization. 2nd edition. World Health Organization. https://apps.who.int/iris/handle/10665/259858

Zhao, Q., De Laender, F. and Van Den Brink, P. J. (2020). Community composition modifies direct and indirect effects of pesticides in freshwater food webs. *Sci Total Environ 739*, 139531. doi:10.1016/j.scitotenv.2020.139531

Zindler, F., Beedgen, F., Brandt, D. et al. (2019). Analysis of tail coiling activity of zebrafish (Danio rerio) embryos allows for the differentiation of neurotoxicants with different modes of action. *Ecotoxicol Environ Saf 186*, 109754. doi:10.1016/j.ecoenv.2019.109754

Conflict of interest

The authors declare not to have any conflict of interest with this manuscript.